Here is the theoretical part of Diffusion Model.

1 Langevin SDE

The Langevin SDE has the following form:

$$X_{t+s} = X_t + \nabla \log p_t(x_t)s + \sqrt{2s}\xi \tag{1}$$

where $X_t \in \mathcal{R}^d, p_t(x_t) = p(X_t = x_t)$, $\xi \sim N(0, I)$, I is identical matrix of $m \times m$. Our goal is to sample from specific $p(x, t)$.

Theorem 1 *The density of Langevin Diffusion Model converges to $p(x)$ over time. In other words, if $X_t \sim p(x)$, then $X_{t+s} \sim p(x)$ for $\forall s > 0$.*

Proof 1 *Let $\mu_t(f) = E\left[f\left(X_t\right)\right]$. Consider $\mu_{t+\tau}(f) = E\left[f\left(X_{t+\tau}\right)\right]$, as $\tau \to 0$. Then*

$$\begin{aligned}
\mu_{t+\tau} =& E\left[f\left(X_t + \nabla \log p_t\left(x_t\right) \cdot \tau + \sqrt{2\tau}\xi\right)\right] \\
=& E\left[f\left(x_t\right) + \nabla^\top f\left(x_t\right)\left(\tau \nabla \log p_t\left(x_t\right) + \sqrt{2\tau}\xi\right)\right. \\
&\left. + \frac{1}{2}\left(\nabla^\top \log p_t(x_t)\tau + \sqrt{2\tau}\xi\right)\nabla^2 f(x_t)\nabla \log p_t(x_t)\tau + \sqrt{2\tau}\xi\right] \\
=& E\left[f\left(x_t\right)\right] + E\left[\tau \nabla^\top f\left(x_t\right) \nabla \log p_t\left(x_t\right)\right] \\
&+ \frac{\tau^2}{2} E\left[\nabla^\top \log p\left(x_t\right) \cdot \nabla^2 f\left(x_t\right) \cdot \nabla \log p\left(x_t\right)\right] + E\left[\tau \xi^\top \nabla^2 f\left(x_t\right) \xi\right]
\end{aligned} \tag{2}$$

The second term:

$$\begin{aligned}
& \tau E\left[\nabla^\top f \nabla \log p_t\right] \\
=& \tau \int \nabla f \cdot \nabla \log p_t p_t dx = \tau \int \nabla f \cdot \nabla p_t dx \\
=& -\tau \int \operatorname{tr}\left(\nabla^2 f\right) \cdot p_t dx = -\tau E\left[\operatorname{tr}\left(\nabla^2 f\right)\right] \\
=& -\tau E\left[\xi^\top \nabla^2 f \xi\right]
\end{aligned} \tag{3}$$

Then

$$\mu_{t+\tau} = E\left[\frac{1}{2}\nabla^\top \log p_t \nabla^2 f \nabla \log p_t\right] \cdot \tau^2 = O\left(\tau^2\right) \tag{4}$$

Hence we have $\frac{d}{dt}(\mu_t) = 0$, i.e. $E[\mu_t] = E[\mu_{t+s}]$ for $\forall s > 0$.

Remark 1 *We define the density of normal distribution $N(x; \mu, \Sigma)$, and its log-density, gradient of density and score as follows:*

$$\begin{cases}
N(x; \mu, \Sigma) = \dfrac{1}{\sqrt{(2\pi)^d |\Sigma|}} e^{-\frac{1}{2}(x-\mu)^\top \Sigma^{-1}(x-\mu)} \\
\log N(x; \mu, \Sigma) = -\dfrac{1}{2}(x-\mu)^\top \Sigma^{-1}(x-\mu) - \log\left(\sqrt{(2\pi)^d |\Sigma|}\right). \\
\nabla_x N(x; \mu, \Sigma) = N(x; \mu, \Sigma)\Sigma^{-1}(x-\mu) \\
\nabla_x \log N(x; \mu, \Sigma) = -\Sigma^{-1}(x-\mu).
\end{cases} \tag{5}$$

Actually, Langevin SDE is not necessary be as above i.e. the diffusion term is not necessary to be $\sqrt{2}$. The reason is to guarantee the stationary distribution of $p_t(x)$. i.e. the term $\frac{\partial p(x,t)}{\partial t} = 0$ in FPK equation. If the diffusion term is $g(t)$, then by FPK equation, we have

$$\nabla_x \cdot (fp - \frac{1}{2}g^2(t)\nabla p) = 0$$

then $f(x,t) = \frac{1}{2}g^2(t)\frac{\nabla_x p(x,t)}{p(x,t)} = \frac{1}{2}g^2(t)\nabla_x \log p(x,t)$.

2 Linear SDE

Then we consider linear SDE having the form:

$$dX_t = (a(t)X_t + b(t))dt + g(t)dW_t \tag{6}$$

By Euler Maruyama method, it can be approximated By

$$\begin{aligned} X_{t+s} &= X_t + (a(t)X_t + b(t))s + g(t)\sqrt{s}\xi \\ &= (1 + a(t)s)X_t + b(t)s + g(t)\sqrt{s}\xi \end{aligned} \tag{7}$$

where $\xi \sim N(0,1)$. Usually we need to consider the expectation, variance and distribution of x. But the stochastic value of x is dependent of x_0. Then first we consider

$$\begin{aligned} E\left[X_{t+s}|X_0\right] - E\left[X_t|X_0\right] &\approx \left(a(t)E\left[X_t|X_t\right] + b(t)s + g(t)\sqrt{s}E[\xi]\right. \\ &= \left(a(t)E\left[X_t|X_0\right] + b(t)\right)s. \end{aligned} \tag{8}$$

Note $e(t) = E\left[X_t|X_0\right]$, then

$$e'(t) = \lim_{s\to 0}\frac{E\left[X_{t+s}|X_0\right] - E\left[X_t|X_0\right]}{s} = a(t)\cdot e(t) + b(t). \quad e(0) = X_0. \tag{9}$$

which is an ODE system, having solution

$$e(t) = \left(X_0 + \int_0^t e^{-\int_0^s a(r)dr} b(s)ds\right)\cdot e^{\int_0^t a(s)ds} \tag{10}$$

Therefore

$$\begin{aligned} E\left[X_t\right] &= E\left[E\left[X_t|X_0\right]\right] = E[e(t)] \\ &= \left(E\left[X_0\right] + \int_0^t e^{-\int_0^s a(r)dr} b(s)ds\right) e^{\int_0^t a(s)ds} \end{aligned} \tag{11}$$

Similarly, Note $\mathrm{Var}\left(X_0|X_0\right) = v(t)$: then $\mathrm{Var}\left(X_{t+s}|X_0\right) = (1+sa(t))^2\,\mathrm{Var}\left(X_t|X_0\right) + sg^2(t)$. Then

$$\begin{aligned} V'(t) &= \lim_{s\to 0}\frac{\mathrm{Var}\left(X_{t+s}|X_0\right) - \mathrm{Var}\left(X_t|X_0\right)}{s} \\ &= \left[\left(a^2(t)s + 2a(t)\right)v(t) + g^2(t)\right]|_{s\to 0} \\ &= 2\alpha(t)V(t) + g^2(t), \qquad V(0) = 0 \end{aligned} \tag{12}$$

Solution is:

$$v(t) = \left(\int_0^t e^{-\int_0^s 2a(r)dr} g^2(s)ds\right) \tag{13}$$

By law of total variance

$$\begin{aligned} \mathrm{Var}\left(X_t\right) =& E\left[X_t^2\right] - E^2\left[X_t\right] = E\left[E\left[X_t^2|X_0\right]\right] - E^2\left[X_t\right] \\ =& E\left[\mathrm{Var}\left(X_t|X_0\right) + E^2\left[X_t|X_0\right]\right] - E^2\left[X_t\right] \\ =& E\left[\mathrm{Var}\left(X_t|X_0\right)\right] + E\left[E^2\left[X_t|X_0\right]\right] - E^2\left[E\left[X_t|X_0\right]\right] \\ =& E\left[\mathrm{Var}\left(X_t|X_0\right)\right] + \mathrm{Var}\left(E\left[X_t|X_0\right)\right) \end{aligned} \tag{14}$$

then

$$\begin{aligned} \mathrm{Var} =& E[V(t)] + \mathrm{Var}(e(t)) \\ =& \left(\int_0^t e^{-\int_0^s 2a(r)dr} g^2(s)ds\right) e^{\int_0^t 2a(s)ds} + e^{\int_0^t 2a(s)ds}\cdot \mathrm{Var}\left(X_0\right). \end{aligned} \tag{15}$$

We have the following theorem which is crucial for diffusion models.

Theorem 2 *If* $X_{t+s} = (1 + a(t)s)X_t + b(t)s + g(t)\sqrt{s}\xi$
then $X_t|X_0 \sim N\left(E\left[X_t|X_0\right], \mathrm{Var}\left(X_t|X_0\right)\right)$*, where* $\left(E\left[X_t|X_0\right] = e(t), \mathrm{Var}\left(X_t|X_0\right) = V(t)\right)$.

It should be noted that $e(t)$ is related to X_0 and t, while $V(t)$ only depends on t!

Next, we will see how the above formula can be applied to diffusion modtels. There are three frameworks to build SDEs for diffusion models, VP, VE and sub-VP.

Definition 1 *Noise function* $\beta(t)$ *. s.t.* $\beta(0) = 0; \beta'(t) \geqslant 0; \beta(t) \to \infty$ *as* $t \to \infty$

2.1 Variance Preserving (VP) SDE

So if we have diffusion model like:

$$\begin{aligned} X_{t_{i+1}} &= \sqrt{1 - \left(\beta\left(t_{i+1}\right) - \beta\left(t_i\right)\right)} X_{t_i} + \sqrt{\left(\beta\left(t_{i+1}\right) - \beta\left(t_i\right)\right)}\xi \\ &= \sqrt{1 - \Delta\beta(t_i)} X_{t_i} + \sqrt{\Delta\beta(t_i)}\xi \end{aligned} \tag{16}$$

Then the conditional distribution is given by:

$$q\left(X_{t_{i+1}}|X_{t_i}\right) = N(x_{t_{i+1}}; \sqrt{1 - \Delta\beta\left(t_i\right)} X_{t_i}, \Delta\beta\left(t_i\right)) \tag{17}$$

Then we need to estimate θ drift term f and diffusion term g:

$$\begin{aligned} f(x,t) &= \lim_{h \to 0} \frac{E\left[X_{t+h} - X_t | X_t = x\right]}{h} \\ &= \lim_{h \to 0} \frac{x\sqrt{1 - \Delta\beta(t)} - x}{h} = -\frac{x}{2}\beta'(t). \\ g(t) &= \sqrt{\lim_{h \to 0} \frac{N\left[X_{t+h}|X_t = x\right]}{h}} = \sqrt{\lim_{h \to 0} \frac{\beta(t+h) - \beta(t)}{h}} = \sqrt{\beta'(t)} \end{aligned} \tag{18}$$

Then the model can be written as $dx = -\frac{x}{2}\beta'(t)dt + \sqrt{\beta'(t)}dW_t$

Then by Theorem 2 we have

$$\begin{cases} E\left[X_t|X_0\right] = X_0 e^{\int_0^t -\frac{1}{2}\beta'(s)ds} = X_0 e^{-\frac{1}{2}\beta(t)} \\ E\left[X_t\right] = E\left[X_0\right] e^{-\frac{1}{2}\beta(t)} \\ V\left(X_t|X_0\right) = \displaystyle\int_0^t e^{\int_0^s \beta'(r)dr}\beta'(s)ds \cdot e^{-\beta(t)} = 1 - e^{-\beta(t)} \\ V\left(X_t\right) = 1 - e^{-\beta(t)} + V\left(X_0\right) e^{-\beta(t)} = 1 + \left(V\left(X_0\right) - 1\right) e^{-\beta(t)}. \end{cases} \tag{19}$$

So as $t \to \infty, \beta(t) \to \infty$, then $E \to 0, V \to 1$, i.e. $X_t|X_0 \sim N\left(E\left[X_t|X_0\right], \mathrm{Var}\left|X_t\right| X_0\right) \to N(0,1)$ as $t \to \infty$.

2.2 Variance-Exploding SDE

Here is the model: $X_{t+h} = X_t + \sqrt{\Delta\beta(t)}\xi$

Similarly we can compute the $f(x,t) \equiv 0$ and $g(t) = \sqrt{\beta(t)}$. Hence

$$\begin{cases} E\left[X_0|X_0\right] = X_0 \\ E\left[X_t\right] = E\left[X_0\right] \\ V\left(X_t|X_0\right) = \displaystyle\int_0^t e^{\int_0^s 0dr}\beta'(s)ds = \beta(t) \\ V\left(X_t\right) = V\left[X_0\right] + \beta(t) \end{cases} \tag{20}$$

So the expectation value is constant and the variance is increasing monotonical.

If we rescale X_t as $Y_t = \frac{X_t}{\sqrt{\beta(t)}}$, then $Y_t \to N(0,1), t \to \infty$.

2.3 Sub-VP SPE

Here, we set the dift and diffusion term as

$$\begin{aligned} f(x,t) &= -\frac{1}{2}\beta'(t) \\ g(t) &= \sqrt{\beta'(t)\left(1-e^{-2\beta(t)}\right)} \end{aligned} \tag{21}$$

As the same, we can compute that.

$$\begin{cases} E\left[X_t|X_0\right] = X_0 e^{-\frac{1}{2}\beta(t)} \\ E\left[X_t\right] = E\left[X_0\right] e^{-\frac{1}{2}\beta(t)} \\ V\left(X_t|X_0\right) = \left(1-e^{-\beta(t)}\right)^2 \\ V\left(X_t\right) = \left(1-e^{-\beta(t)}\right)^2 + V\left(X_t\right) e^{-\beta(t)}. \end{cases} \tag{22}$$

We can find out that the variance is always smaller thar VP SDE.

Remark 2 *To sum up, finally we hope that X_t converges to a normal distribution by choosing different drift and diffusion functions. For generative model, the goal is to sample from a Data distribution p_{data}. We have known that if we set the initial distribution $p_0(x_0) = p(X_0 = x_0) \sim p_{data}$, then after $t = T$, the distribution of X_t is tend to be $N(0,1)$ under certain conditions.*

So the idea is backward: if we sample from $X_T \sim N(0,1)$, and then run SDE backwards, could we get the initial distribution?

3 Reverse SDE

Assume we have forward SDE: from $X_0 \sim p_0, X_T \sim p_T$,

$$dX_t = f(X_t,t)dt + g(t)dW_t \tag{23}$$

Then we define the reverse SDE as: from $X_T \sim p_T$,

$$d\bar{X}_t = \bar{f}(\bar{X}_t,t)dt + \bar{g}(t)d\bar{W}_t \tag{24}$$

where $\bar{W}_t$ is Brownian Motion runns backward in time, i.e. $\bar{W}_{t-s} - \bar{W}_t$ is independent of $\bar{W}_t$. We can approximate by EM:

$$\bar{X}_{t-s} - \bar{X}_t = -s\bar{f}(\bar{X}_t,t) + \sqrt{s}\bar{g}(t)\xi \tag{25}$$

So the problem is: If given f, g, are there $\bar{f}, \bar{g}$ s.t. the reverse time diffusion process $\bar{X}_t$ has the same distribution as the forward process X_t? Yes!

Theorem 3 *The reverse SDE with $\bar{f}, \bar{g}$ having the following form has the same distribution as the forward SDE 23:*

$$\begin{cases} \bar{f}(x,t) = f(x,t) - g^2(x,t)\dfrac{\partial}{\partial x}\log p_t(x) \\ \bar{g} = g(t) \end{cases} \tag{26}$$

i.e.

$$d\bar{X}_t = \left[f(\bar{X}_t,t) - g^2(t)\log p_t(x_t)\right]dt + g(t)d\bar{W}_t \tag{27}$$

Proof 2 *The proof is skipped.*

This theroem allows us to learn how to generate samples from p_{data}.

Algorithm 1 *:*

Step1. Select $f(x,t)$ and $g(t)$ with affine drift coefficients s.t. $X_T \sim N(0,1)$

Step2. Train a network $s_\theta(x,t) = \frac{\partial}{\partial x}\log p_t(x)$ where $p_t(x) = p(X_t = x)$ is the forward distribution.

Step3. Sample X_T from $N(0,1)$, then run reverse SDE from T to 0:

$$\bar{X}_{t-s} = \bar{X}_t + s\left[g^2(t)s_\theta(\bar{X}_t,t) - f(\bar{X}_t,t)\right] + \sqrt{s}g(t)\xi \tag{28}$$

4 Loss function

Normally we can define the loss function as follows:

$$\begin{aligned} L_\theta &= \frac{1}{T}\int_0^T \lambda(t) \underset{x_0\sim p_{data}}{E}\left[\underset{x_t\sim p_{t|0}(x_t|x_0)}{E}\left[\|s_\theta(x_t,t) - \nabla_{x_t}\log p_t(x_t)\|^2\right]\right] dt \\ &= \underset{t\sim U(0,T)}{E}\left[\lambda(t)\underset{x_0\sim p_{data}}{E}\left[\underset{x_t\sim p_{t|0}(x_t|x_0)}{E}\left[\|s_\theta(x_t,t) - \nabla_{x_t}\log p_t(x_t)\|^2\right]\right]\right] \end{aligned} \tag{29}$$

It should be clearified that $p_{t|0}(x_t|x_0) = p(X_t = x_t|X_0 = x_0)$. So

$$p_t(x_t) = \int p_{t|0}(x_t|x_0)p_0(x_0)dx_0 = E_{x_0\sim p_{data}}\left[p_{t|0}(x_t|x_0)\right]$$

. Then $p_t(x) = p\left(X_t = x\right)$. $p_{t|0}(x|y) = p\left(X_t = x|x_0 = y\right)$.

$$\begin{aligned} \nabla\log p_t(x) &= \frac{1}{p_t(x)}\nabla p_t(x). \\ &= \frac{1}{p_t(x)}\nabla\int p_{t|0}(x|y)p_0(y)dy \\ &= \frac{1}{p_t(x)}\int \nabla p_{t|0}(x|y)p_0(y)dy \\ &= \frac{1}{p_t(x)}\int \frac{\nabla p_{t|0}(x|y)}{p_{t|0}(x|y)}p_0(y)\cdot p_{t|0}(x|y)dy \\ &= \int \nabla_x\log\left(p_{t|0}(x|y)\right)\cdot p_{0|t}(y|x)dy \\ &= \underset{y\sim p_{0|t}(y|x)}{E}\left[\nabla_x\log\left(p_{t|0}(x|y)\right)\right] \end{aligned} \tag{30}$$

Where we have used the following lemma:

Lemma 1

$$\underset{x_0\sim p_0}{E}\left[\underset{x_t\sim p_{t|0}(\cdot|x_0)}{E}\left[\underset{x_0'\sim p_{0|t}(\cdot|x_t)}{E}\left[f(x_t,x_0')\right]\right]\right] = \underset{x_0\sim p_0}{E}\left[\underset{x_t\sim p_{t|0}(\cdot|x_0)}{E}\left[f(x_t,x_0)\right]\right] \tag{31}$$

Proof 3 *Easy to prove.*

Then we can rewrite the loss function as:

$$\begin{aligned} L_\theta &= \underset{t\sim U(0,T)}{E}\left[\lambda(t)\underset{x_0\sim p_{data}}{E}\left[\underset{x_t\sim p_{t|0}(x_t|x_0)}{E}\left[\left\|S_\theta\left(x_t,t\right) - \frac{\partial}{\partial x_t}\log p_t\left(x_t\right)\right\|^2\right]\right] \\ &\leqslant \underset{t\sim U(0,T)}{E}\left[\lambda(t)\underset{x_0\sim p_{data}}{E}\left[\underset{x_t\sim p_{t|0}(x_t|x_0)}{E}\left[\underset{y\sim p_{data}}{E}\left[\left\|S_\theta\left(x_t,t\right) - \nabla_{x_t}\log\left(p_{t|0}(x_t|y)\right)\right\|^2\right]\right]\right]\right] \\ &= \underset{t\sim U(0,T)}{E}\left[\lambda(t)\underset{x_0\sim p_{data}}{E}\left[\underset{x_t\sim p_{t|0}(x_t|x_0)}{E}\left[\left\|S_\theta\left(x_t,t\right) - \nabla_{x_t}\log\left(p_{t|0}\left(x_t|x_0\right)\right\|^2\right]\right]\right] \end{aligned} \tag{32}$$

Since $p_{t|0}\left(x_t|x_0\right) = p\left(X_t = x_t|X_0 = x_0\right)$ has been discussed:

$$p_{t|0}\left(x_t|x_0\right) \sim N\left(x_t; E\left[X_t = x_t|X_0 = x_0\right], \mathrm{Var}\left(X_t = x_t|X_0 = x_0\right)\right).$$

Then by theorem 2, x can be written as $x = e(t, X_0) + \sqrt{V(t)}\xi$, where $\xi \sim N(0,1)$, then the score function is:

$$\frac{\partial}{\partial x} \log p_{t|0}(x|x_0) = -\frac{x - E_{t|0}[x|x_0]}{\mathrm{Var}_{t|0}(x|x_0)} = -\frac{x - e(t, X_0)}{V(t)} \sim -N\left(0, \frac{1}{V(t)}\right) \tag{33}$$

So

$$\begin{aligned} L_\theta &= \underset{t\sim U(0,T)}{E}\left[\lambda(t) \underset{x_0\sim p_{data}}{E}\left[\underset{\xi\sim N(0,1)}{E}\left[\left\| s_\theta\left(\sqrt{V(t)}\xi + e(t, X_0), t\right) + \frac{\xi}{\sqrt{V(t)}}\right\|^2\right]\right]\right] \\ &= \underset{t\sim U(0,T)}{E}\left[\lambda(t) \underset{x_0\sim p_{data}}{E}\left[\frac{1}{V(t)} \underset{\xi\sim N(0,1)}{E}\left[\left\| \xi_\theta\left(\sqrt{V(t)}\xi + e(t, X_0), t\right) - \xi\right\|^2\right]\right]\right] \end{aligned} \tag{34}$$

where $\xi_\theta = -\sqrt{V(t)}s_\theta$ is called denoising network.

5 With Classifier

Though we can produce pictures by sampling from normal distribution, we still cannot control what we will generate. What we want to do is something like: "Give me the pictures of number 6", then the model can sample from the normal distribution and do the denoising to generate pics of 6.

Usually, we can do something like: train a model for every class label. This do make the model smaller, but increases number of models. Think about it, when the label is TEXT, it is impossiable to train a model for each sentences.

So, the initial distribution is $p_0(x|y)$ given the label y. Similarly, we will convert the data distribution $p_{data}(x|y)$ to final distribution, normal distribution expected. Then we SDE becomes: $X_t \sim p_t(x|y)$

$$\begin{aligned} P_t(x \mid y) = P(X_t = x \mid y) &= \frac{P(y \mid X_t = x) P(X_t = x)}{P(y)} \\ \Rightarrow \log(P_t(x \mid y)) &= \log(P(y \mid X_t = x)) + \log(P(X_t = x)) - \log(P(y)) \\ \Rightarrow \nabla_x \log(P_t(x \mid y)) &= \nabla_x \log(P(y \mid X_t = x)) + \nabla_x \log(P(X_t = x)) \end{aligned} \tag{35}$$

We have finished training $\nabla_x \log(p(X_t = x))$ in sampling. Then we need to estimate $\nabla_x \log(p(y \mid X_t = x))$. This is the conditional protability, we end up with a sharp factor s: $P'(y \mid X_t = x)$, then:

$$\nabla_x \log(p_t(x \mid y)) = S\nabla_x \log(p(y \mid x_t = x)) + \nabla_x \log(p(x_t = x)) \tag{36}$$

Note $\omega_\theta(y \mid x, t)$ to learn $s\nabla_x \log(p(y \mid X_t = x))$